Chapter 2

Simple Randomized Experiments

2.1 Why Randomization?

Suppose we observe outcomes $(Y_1, ..., Y_n)$ for n subjects, each of whom are either treated $(A = 1)$ or not $(A = 0)$, and we want to learn the effect of the treatment A on outcome Y, say on average. An initial idea might be to compare the average outcome for those who receive treatment versus control:

$$\sum_{i:A_i=1} Y_i \quad \text{versus} \quad \sum_{i:A_i=0} Y_i$$

However, as discussed previously, any differences we see could be spurious, i.e., explained by something else. For example, high-tech NICUs have higher mortality rates than low-tech NICUs, but this is because high-tech NICUs see the sickest infants. In general, differences in outcomes might just be due to the people receiving treatment being inherently different from those receiving control.

A second option might be to try to measure any and all variables X that could explain any differences in outcomes, and then do a stratified (adjusted) analysis that only compares subjects with the same X values. In the NICU example, one might try to measure every possible facet of babies' health, such as birthweight, gestational age, family medical history, mother's smoking status, relevant biomarker values and lab results, and so on. There are at least three severe difficulties with this approach:

1. We often simply do not know every single $X = (X_1, X_2, ..., X_{1000}, ...)$ that could explain *any* differences in outcomes.

2. Even if we did know every single possible X with certainty, it might be impossible or too expensive to measure every single one of them.

3. And even if we could measure every single X, there may be so many that the curse of dimensionality would make estimation impossibly difficult (e.g., few if any subjects would have the same or similar Xs in every dimension).

So is assumption-free causal inference hopeless? Luckily not; it turns out there is a simple yet beautiful solution if we can control who gets treatment: assign it randomly! For example, one could flip a coin to decide whether each subject gets treatment versus control. Surprisingly, the benefits of randomization were largely unknown until relatively recently in the long history of science (according to the OED its first recorded use was due to R.A. Fisher in 1926).

Why does random treatment assignment work? Randomization ensures that the treatment is completely independent of *all* subject characteristics, whether measured or not. In other words, the treated look *exactly the same* as the untreated, in expectation, and not only for all measured variables X but also for *any* unmeasured variables U. Thus any observed differences in the outcomes for the treated versus untreated must be due to the treatment, since it is the only systematic way in which the groups differ.

Another way to think about why randomization works is in terms of potential outcomes. Suppose each subject has two potential outcomes, Y^1 and Y^0, with the former revealed by treatment and the latter revealed by control, so that $Y = AY^1 + (1-A)Y^0$. Then, by randomly assigning treatment A, we are taking two random samples – one of the Y^1 values and another of the Y^0 values. Random samples yield unbiased estimators of population means, so the average outcomes in the two groups will be unbiased estimates of the corresponding average potential outcomes.

Of course, we can also prove randomization works mathematically:

Proposition 2.1. *Let* $(A, Y) \sim \mathbb{P}$ *and assume:*

1. *Consistency:* $Y = Y^a$ *whenever* $A = a$.
2. *Randomization:* $A \perp\!\!\!\perp Y^a$ *for each* a.

Then

$$\mathbb{E}(Y \mid A = a) = \mathbb{E}(Y^a).$$

Proof. It follows that

$$\mathbb{E}(Y \mid A = a) = \mathbb{E}(Y^a \mid A = a) = \mathbb{E}(Y^a)$$

using consistency in the first equality and randomization in the second. □

Remark 2.1. Make sure not to confuse $A \perp\!\!\!\perp Y^a$ with $A \perp\!\!\!\perp Y$: these are very different. $A \perp\!\!\!\perp Y^a$ means treatment is independent of potential outcomes (which can be viewed as "pre-treatment" variables that exist just prior to the treatment assignment), and reflects that treatment is not confounded; $A \perp\!\!\!\perp Y$ means treatment is independent of the *observed* outcome, and would for example be a consequence of treatment not only being unconfounded but also ineffective (e.g., $Y^1 = Y^0$). Always remember to distinguish potential outcomes from observed outcomes.

Remark 2.2. Although Proposition 1 gives an identification result for the mean potential outcome, its assumptions are sufficient for identifying the entire distribution of potential outcomes as $\mathbb{P}(Y^a \leq t) = \mathbb{P}(Y \leq t \mid A = a)$.

Proposition 1 shows that treatment assignment need not necessarily be a subject-specific coin flip – for the purposes of achieving identification of the potential outcome distribution, treatment just needs to be independent of potential outcomes. This leads to the following definition of a randomized experiment:

Definition 2.1. A study is a randomized experiment if the treatment assignment is both *probabilistic* and *known*.

There are many types of experimental designs. For example, letting $A^n = (A_1, ..., A_n)$:

- Bernoulli: Treatments assigned via independent coin flips, i.e., $\mathbb{P}(A^n = a^n) = (1/2)^n$ for every $a^n = (a_1, ..., a_n) \in \{0, 1\}^n$.
- Stratified Bernoulli: Treatments assigned via independent *biased* coin flips depending on covariates, i.e., $\mathbb{P}(A^n = a^n \mid X^n) = \prod_i \mathbb{P}(A_i = a_i \mid X_i)$.
- Completely randomized: n_1 of n subjects randomly assigned to treatment, i.e., $\mathbb{P}(A^n = a^n) = 1/\binom{n}{n_1}$ for $\sum_i a_i = n_1$.
- Matched pairs: Matched pairs are constructed and one is treated in each, i.e., $\mathbb{P}(A^n = a^n \mid X^n) = 1/2^{n/2}$.

One design may be favored over another due to efficiency or feasibility, for example.

2.2 Testing: Fisher's Sharp Null

Jerzy Neyman was the first to introduce potential outcomes (in 1923), but R.A. Fisher was the first to really advocate for randomization (in 1925).

Fisher was interested in testing the *sharp null hypothesis*

$$H_0 : Y_i^1 = Y_i^0 \text{ for all } i$$

which says that treatment has no effect whatsoever – not only is the mean of Y^1 exactly equal to that of Y^0, but the distributions are equal and further each individual potential outcome is exactly the same under both treatment and control. This is a strong null with lots of structure, in line with Fisher's perspective that one should "make your theories elaborate".

Recall to test a generic null hypothesis H_0 we need (1) a statistic T, and (2) its distribution under the null. Then one can obtain the infamous statistic known as a p-value, i.e., $\mathbb{P}_{H_0}(T \geq t_{obs})$, the chance under the null of seeing data as extreme as that which was actually observed.

To test Fisher's sharp null, we can use as a statistic any summary measure of how treatment changes outcomes; for example, a simple yet common choice is the absolute difference-in-means

$$T(A^n, Y^n) = \left| \frac{1}{n_1} \sum_{i:A_i=1} Y_i - \frac{1}{n_0} \sum_{i:A_i=0} Y_i \right| = \left| \frac{\mathbb{P}_n(AY)}{\mathbb{P}_n(A)} - \frac{\mathbb{P}_n\{(1-A)Y\}}{\mathbb{P}_n(1-A)} \right|$$

Note this test statistic will be large if the treated versus untreated means differ, but not if the treatment only changes non-central aspects of the distribution, e.g., the variance.

Armed with a test statistic, we now need to know its distribution under the null. This is actually easy, and typically part of the motivation for using the sharp null: it yields tractable null distributions, which can be computed in a non-asymptotic and distribution-free manner. To illustrate, consider the following completely randomized experiment, simulated in R:

```
> set.seed(100)
> ## simulate fake data
> n <- 10; a <- rep(c(1,0),5); y <- a*rnorm(n,1)+(1-a)*rnorm(n,-1)
> cbind(a,y)
      a          y
 [1,] 1  0.4978076
 [2,] 0 -0.9037255
 [3,] 1  0.9210829
 [4,] 0 -0.2601595
 [5,] 1  1.1169713
 [6,] 0 -1.0293167
 [7,] 1  0.4182093
 [8,] 0 -0.4891437
 [9,] 1  0.1747406
[10,] 0  1.3102968
>
> ## compute test statistic
> (tobs <- abs(mean(y[a==1])-mean(y[a==0])))
[1] 0.9001721
```

Here the observed value of the difference-in-means test statistic is $T(A^n, Y^n) \approx 0.9$. We can also compute the value of this statistic under the null, for any randomization, since under the null the potential outcomes are exactly the same, i.e., $Y^0 = Y^1 = Y$. Therefore we can obtain the null distribution of T by permuting the A^n vector (according to the known treatment assignment mechanism), while keeping the Y^n vector fixed, computing the corresponding value of the test statistic T, which yields the corresponding distribution $\mathbb{P}_{H_0}(T \leq t)$. A p-value can be computed by simply counting the proportion of permutations with test statistics larger than that which was observed.

This can be accomplished in R with:

```
> ## permute treatments to simulate null
> t <- NULL; for (j in 1:10000){
+     asim <- sample(a)
+     t <- c(t, abs(mean(y[asim==1])-mean(y[asim==0]))) }
>
> ## compute p-value
> mean(t>=tobs)
[1] 0.0837
```

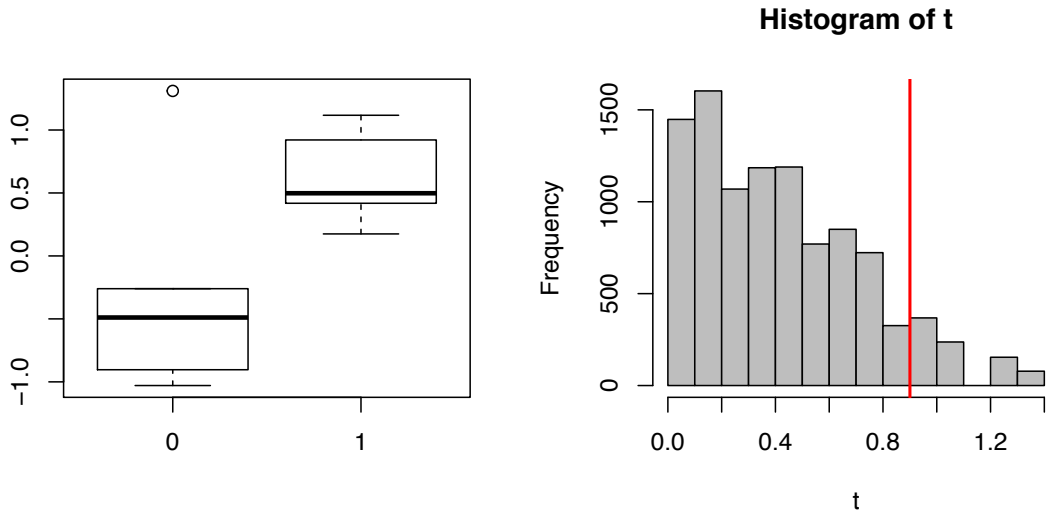

Figure 2.1: Boxplot of simulated data (left), and histogram of permutation-based null distribution with the red vertical line denoting the observed test statistic value (right).

The data and results are shown in Figure 2.1. In this simulation, the p-value is 0.084 and so there is sufficient evidence to reject the sharp null hypothesis of no individual treatment effect at level $\alpha = 0.10$.

Mathematically, for a completely randomized experiment where a fixed number n_1 are treated, the null distribution can be written as

$$\mathbb{P}_{H_0}(T \geq t) = \mathbb{P}_{H_0}\{T(A^n, y^n) \geq t\} = \sum_{a^n \in \mathcal{A}} \mathbb{1}\{T(a^n, y^n) \geq t\}\mathbb{P}(A^n = a^n)$$
$$= \sum_{a^n : \sum_i a_i = n_1} \frac{\mathbb{1}\{T(a^n, y^n) \geq t\}}{\binom{n}{n_1}}$$

In theory we can compute this distribution exactly; in practice if n is large we may need to resort to simulation (e.g., sample K of the $\binom{n}{n_1}$ randomizations). However the distribution can be simulated with arbitrarily high accuracy by taking K large enough.

Remark 2.3. The null distribution calculation above treats the (potential) outcomes y^n as fixed; this can be viewed as an assumption that Y^a is not a random variable, or the probability can just be defined conditionally given the random potential outcomes.

Fisher's permutation-style test is simple but impressive: it gives an exact distribution-free p-value for testing H_0, which is valid for any n. Nonetheless here are some caveats:

- The power of the test depends heavily on the choice of statistic, e.g., the difference-in-means test statistic will have no power against a treatment that makes outcomes bimodal or otherwise more variable.
- Fisher's test is of the *sharp null* of no individual effect, not of no average effect – in fact rejecting Fisher's null could still mean there is no effect on average.

2.3 Estimation: Sample Average Effects

In the 1920s and 1930s, Fisher and Neyman had some heated debates about whether testing Fisher's sharp null should be the primary goal or not; in contrast to Fisher, Neyman advocated for estimation rather than testing, and focused on average effects. Average effects might be considered more relevant for policy decisions, since they indicate how a population would fare on average if all versus none were treated; in contrast rejecting the sharp null only indicates that treatment has *some* effect, without saying much about what kind.

The *sample average treatment effect* is given by

$$\psi_n = \frac{1}{n}\sum_{i=1}^{n}(y_i^1 - y_i^0)$$

This parameter is different from those we will study later in that it is a functional of the particular sample rather than of a population distribution (i.e., strictly speaking it is a data-dependent parameter, which is why we index it with n).

Remark 2.4. In this section we treat potential outcomes as fixed, not random; this is equivalent to treating probability statements as conditional on the potential outcomes. Note however that even if the potential outcomes are fixed, the observed outcome is random since it is a function of the random treatment: $Y = Ay^1 + (1-A)y^0$.

A natural estimator for ψ_n is the difference-in-means

$$\widehat{\psi} = \frac{1}{n_1}\sum_{i:A_i=1} Y_i - \frac{1}{n_0}\sum_{i:A_i=0} Y_i = \frac{\mathbb{P}_n(AY)}{\mathbb{P}_n(A)} - \frac{\mathbb{P}_n\{(1-A)Y\}}{\mathbb{P}_n(1-A)}$$

We will now characterize the bias and variance of this estimator and discuss inference.

Proposition 2.2. *The difference-in-means estimator is unbiased for ψ_n in a completely randomized experiment, assuming consistency (i.e., $Y = Ay^1 + (1-A)y^0$).*

Proof. By definition, in a completely randomized experiment, we have

$$\begin{aligned}\mathbb{P}(A_1 = 1) &= \sum_{\sum_{i>1} a_i = n_1 - 1} \mathbb{P}(A_1 = 1, A_2 = a_2, ..., A_n = a_n) \\ &= \sum_{\sum_{i>1} a_i = n_1 - 1} \binom{n}{n_1}^{-1} = \frac{\binom{n-1}{n_1-1}}{\binom{n}{n_1}} = \frac{n_1}{n}\end{aligned}$$

and similarly for all other $i > 1$. Therefore

$$\begin{aligned}\mathbb{E}(\widehat{\psi}) &= \mathbb{E}\left\{\frac{1}{n_1}\sum_{i=1}^{n} A_i y_i^1 - \frac{1}{n_0}\sum_{i=1}^{n}(1 - A_i)y_i^0\right\} \\ &= \frac{1}{n_1}\sum_{i=1}^{n}\mathbb{E}(A_i)y_i^1 - \frac{1}{n_0}\sum_{i=1}^{n}\{1 - \mathbb{E}(A_i)\}y_i^0 = \frac{1}{n}\sum_{i=1}^{n}(y_i^1 - y_i^0)\end{aligned}$$

where the first equality follows by consistency, and the last since $\mathbb{P}(A_i = 1) = n_1/n$. □

As mentioned previously, the intuition behind unbiasedness in this setup is that the treatments pick out random samples of Y^1 and Y^0 potential outcomes.

Now we will explore the variance of $\widehat{\psi}$, which is critical for confidence intervals and hypothesis tests; its calculation requires some care since the A_is are not independent (e.g., in the $n = 2$ case, if $A_1 = 1$ then it must be the case that $A_2 = 0$).

Proposition 2.3. *For a completely randomized experiment, and assuming consistency, the variance of the difference-in-means estimator is given by*

$$var(\widehat{\psi}) = \frac{\sigma_n^2(y^1)}{n_1} + \frac{\sigma_n^2(y^0)}{n_0} - \frac{\sigma_n^2(y^1 - y^0)}{n} \tag{2.1}$$

where

$$\sigma_n^2(v) = \frac{1}{n-1}\sum_{i=1}^{n}\left(v_i - \frac{1}{n}\sum_{j=1}^{n} v_j\right)^2$$

denotes the finite sample variance of $(v_1, ..., v_n)$.

Proof. See the appendix of Chapter 6 in Imbens and Rubin [2015] □

A finite-sample central limit theorem implies under some regularity conditions that

$$\frac{\widehat{\psi} - \psi_n}{\sqrt{\text{var}(\widehat{\psi})}} \rightsquigarrow N(0, 1)$$

Therefore to construct large-sample confidence intervals, one needs to estimate the variance $\text{var}(\widehat{\psi})$ in (2.1). The first two terms in this variance can be estimated with

$$\widehat{\sigma}_n^2(y^a) = \frac{1}{n_a - 1} \sum_{i:A_i=a} \left(Y_i - \frac{1}{n_a} \sum_{j:A_j=a} Y_i \right)^2$$

but the third term is the finite-sample variance of the treatment effects $(y_i^1 - y_i^0)$, and involves product terms like $y_i^1 y_i^0$ which can never be observed together. Thus the third term cannot be consistently estimated; however it can be upper bounded as, for example

$$\text{var}(\widehat{\psi}) \leq \frac{\sigma_n^2(y^1)}{n_1} + \frac{\sigma_n^2(y^0)}{n_0} \tag{2.2}$$

which will yield conservative (at worst) inference, when used to construct confidence intervals. Tighter bounds can be achieved with the Cauchy-Schwarz inequality or Frechet-Hoeffding bounds [Aronow et al., 2014].

2.4 Population Average Effects

In this section we move to population rather than finite-sample effects. These effects are useful since in many cases one views the data as a haphazard sample from some larger population that is really of interest.

Suppose we observe an iid sample $(Z_1, ..., Z_n) \sim \mathbb{P}$ with $Z = (A, Y)$. In this section our goal is to estimate the population average effect

$$\psi = \mathbb{E}(Y^1 - Y^0) = \mathbb{E}(Y \mid A = 1) - \mathbb{E}(Y \mid A = 0)$$

rather than the sample average effect ψ_n from before. We rely on the following two assumptions:

1. Consistency: $Y = Y^a$ if $A = a$.
2. Bernoulli randomization: $A \perp\!\!\!\perp Y^a$ with $\mathbb{P}(A = 1) = \pi$.
3. Finite variance: Y has finite conditional variance given $A = a$.

Remark 2.5. In a Bernoulli trial, the number treated $N_1 = \sum_i A_i \sim \text{Bin}(n, \pi)$ is random, not fixed. In this section we also view the potential outcomes as random, not fixed.

Recall the difference-in-means estimator is given by

$$\widehat{\psi} = \frac{1}{\sum_i A_i} \sum_{i:A_i=1} Y_i - \frac{1}{\sum_i (1 - A_i)} \sum_{i:A_i=0} Y_i = \frac{\mathbb{P}_n(AY)}{\mathbb{P}_n(A)} - \frac{\mathbb{P}_n\{(1 - A)Y\}}{\mathbb{P}_n(1 - A)}$$

Now we will study the properties of $\widehat{\psi}$: bias, variance, and limiting distribution. We will see that very precise estimation and inference are possible for the causal effect ψ in Bernoulli trials, under essentially no assumptions beyond consistency.

2.4.1 Properties of the Difference-in-Means Estimator

Theorem 2.1. *Assume consistency. In a Bernoulli trial, the difference-in-means estimator is unbiased for ψ and has variance no greater than*

$$\frac{2}{(n+1)}\left(\frac{\sigma_1^2}{\pi}+\frac{\sigma_0^2}{1-\pi}\right)$$

where $\sigma_a^2 = var(Y \mid A=a)$.

Proof. Let $\widehat{\pi} = \mathbb{P}_n(A)$ and just consider the first term $\widehat{\mu}_1 = \mathbb{P}_n(AY)/\widehat{\pi}$. We have

$$\begin{aligned}\mathbb{E}(\widehat{\mu}_1 \mid A^n) &= \frac{1}{\widehat{\pi}}\mathbb{E}\Big\{\mathbb{P}_n(AY) \mid A^n\Big\} = \frac{1}{\widehat{\pi}}\mathbb{P}_n\Big\{A\mathbb{E}(Y \mid A^n)\Big\} \\ &= \frac{1}{\widehat{\pi}}\mathbb{P}_n\Big\{A\mathbb{E}(Y \mid A=1)\Big\} = (\widehat{\pi}\mu_1)/\widehat{\pi} = \mu_1\end{aligned}$$

where the third equality used the iid assumption. Unbiasedness now follows by iterated expectation, and consistency follows from the weak law of large numbers and continuous mapping theorem. The logic is exactly the same for $\widehat{\mu}_0 = \mathbb{P}_n\{(1-A)Y\}/(1-\widehat{\pi})$.

By the law of total variance we have

$$\text{var}(\widehat{\mu}_1) = \text{var}\Big\{\mathbb{E}(\widehat{\mu}_1 \mid A^n)\Big\} + \mathbb{E}\Big\{\text{var}(\widehat{\mu}_1 \mid A^n)\Big\}$$

Note $\text{var}\{\mathbb{E}(\widehat{\mu}_1 \mid A^n)\} = \text{var}(\mu_1) = 0$ from above, and

$$\begin{aligned}\text{var}(\widehat{\mu}_1 \mid A^n) &= \left(\frac{1}{n\widehat{\pi}}\right)^2 \sum_{i=1}^n A_i \text{var}\left(Y_i \mid A^n\right) \\ &= \left(\frac{1}{n\widehat{\pi}}\right)^2 \sum_{i=1}^n A_i\sigma_1^2 = \frac{\sigma_1^2}{N_1}\mathbb{1}(N_1>0)\end{aligned}$$

where we used independence and defined $\sigma_1^2 = \text{var}(Y \mid A=1)$ and $N_1 = n\widehat{\pi}$. Now

$$\text{var}(\widehat{\mu}_1) = \mathbb{E}\Big\{\text{var}(\widehat{\mu}_1 \mid A^n)\Big\} \leq \frac{2\sigma_1^2}{(n+1)\pi}$$

by the expected binomial reciprocal result (Lemma A.2) of Devroye et al. [1996]. The same logic applies to $\widehat{\mu}_0$, and iterated expectation shows that the covariance term $\text{cov}(\widehat{\mu}_1, \widehat{\mu}_0)$ is exactly zero, which gives the result. □

Theorem 2.2. *Assume consistency. For a Bernoulli trial, the difference-in-means estimator is root-n consistent and asymptotically normal with*

$$\sqrt{n}(\widehat{\psi}-\psi) \rightsquigarrow N\left(0,\ \frac{\sigma_1^2}{\pi}+\frac{\sigma_0^2}{1-\pi}\right)$$

where $\sigma_a^2 = var(Y \mid A=a)$.

Proof. We again focus on μ_1 and its estimator. Note we have

$$\begin{aligned}\widehat{\mu}_1 - \mu_1 &= \frac{\mathbb{P}_n(AY)}{\widehat{\pi}} - \mu_1 = \mathbb{P}_n\left\{\frac{A}{\widehat{\pi}}(Y-\mu_1)\right\} \\ &= \mathbb{P}_n\left\{\frac{A}{\pi}(Y-\mu_1)\right\} + \left(\frac{1}{\widehat{\pi}} - \frac{1}{\pi}\right)\mathbb{P}_n\left\{A(Y-\mu_1)\right\} \\ &= \mathbb{P}_n\left\{\frac{A}{\pi}(Y-\mu_1)\right\} + O_{\mathbb{P}}(1/\sqrt{n})O_{\mathbb{P}}(1/\sqrt{n})\end{aligned}$$

where the last equality follows by the central limit theorem, which implies $\sqrt{n}\{\mathbb{P}_n(V) - \mathbb{E}(V)\} = O_{\mathbb{P}}(1)$ for any iid V with finite mean and variance, together with the fact that $(\widehat{\pi}, \pi)$ are bounded away from zero.

Therefore

$$\widehat{\mu}_1 - \mu_1 = \mathbb{P}_n\left\{\frac{A}{\pi}(Y-\mu_1)\right\} + o_{\mathbb{P}}(1/\sqrt{n})$$

since $O_{\mathbb{P}}(1/\sqrt{n})O_{\mathbb{P}}(1/\sqrt{n}) = O_{\mathbb{P}}(1/n) = o_{\mathbb{P}}(1/\sqrt{n})$, which from the CLT gives

$$\sqrt{n}\Big(\widehat{\mu}_1 - \mu_1\Big) \rightsquigarrow N\left(0, \operatorname{var}\left\{\frac{A}{\pi}(Y-\mu_1)\right\}\right)$$

The logic for the $\widehat{\mu}_0$ part is analogous.

□

Theorem 2.1 is powerful in showing that mean counterfactuals can be estimated very precisely – with zero bias and variance that scales like $1/n$ – in Bernoulli trials, using no assumptions other than consistency and finite variance. Randomization allows accurate and essentially assumption-free causal inference!

Similarly, Theorems 2.1 and 2.2 also pave the way for inference, in the form of confidence intervals and hypothesis tests. Namely, finite sample confidence intervals could be constructed based on Theorem 2.1 using bounds on the conditional variances σ_a^2, and Theorem 2.2 implies for example that an asymptotic 95% CI is given by

$$\widehat{\psi} \pm \left(\frac{1.96}{\sqrt{n}}\right)\widehat{\operatorname{sd}}\left\{\frac{A(Y-\widehat{\mu}_1)}{\pi} - \frac{(1-A)(Y-\widehat{\mu}_1)}{1-\pi}\right\}.$$

Remark 2.6. We saw above that the asymptotic variance of the difference-in-means estimator in a Bernoulli experiment is given by

$$\frac{\sigma_1^2}{\pi} + \frac{\sigma_0^2}{1-\pi}.$$

One interesting thing to note about this variance comes the perspective of experimental design: what is the best choice of π for optimizing efficiency? In fact, it is easy to show that

$$\arg\min_{\pi}\left(\frac{\sigma_1^2}{\pi} + \frac{\sigma_0^2}{1-\pi}\right) = \frac{\sigma_1}{\sigma_0 + \sigma_1}$$

so for optimal efficiency the proportion treated should match the standard deviation of treated outcomes, as a fraction of the total standard deviation for treated and untreated outcomes. This is intuitive – if outcomes are more variable among treated patients than controls (i.e., $\sigma_1 > \sigma_0$) then more patients should be assigned to treatment to counterbalance the extra noise.

2.4.2 Sample versus Population Effects

Here we point out an interesting connection between sample effect estimation in completely randomized experiments and population effect estimation in Bernoulli experiments.

Based on Theorem 2.2, an asymptotic 95% CI for ψ in a Bernoulli experiment is given by

$$\widehat{\psi} \pm 1.96\sqrt{\frac{\widehat{\sigma}_1^2}{n\widehat{\pi}} + \frac{\widehat{\sigma}_0^2}{n(1-\widehat{\pi})}}$$

where $\widehat{\sigma}_a^2 \equiv \sigma_n^2(y^a)$ is the usual sample variance among the treated ($a = 1$) and controls ($a = 0$), which we used in our analysis of the difference-in-means as an estimator of the *sample* average effect in completely randomized experiments (e.g., Proposition 2.3).

In fact, $\widehat{\psi}$ is the exact same point estimate of the sample effect that we analyzed in completely randomized experiments, and similarly the exact same confidence interval

$$\widehat{\psi} \pm 1.96\sqrt{\frac{\widehat{\sigma}_1^2}{n\widehat{\pi}} + \frac{\widehat{\sigma}_0^2}{n(1-\widehat{\pi})}}$$

is also valid (possibly conservative) in completely randomized experiments, guaranteeing at least 95% coverage of the sample effect. (This results from using the naive bound that $\sigma_n^2(y^1 - y^0) \geq 0$ as in (2.2)).

Thus, not only is the estimator for the population effect exactly the same as that for the sample effect, but confidence intervals for the population effect are also valid for the sample effect, being at worst conservative. This is an archetypal example of how finite-sample and population-based frameworks can coincide.

Note that, although population-based confidence intervals are valid for sample effects, the converse is not necessarily true: it is easier to estimate sample effects, in the sense that the same estimators have smaller variances relative to sample versus population effects. Thus a confidence interval for a sample effect may not be valid for a population effect. For example, Imbens [2004] shows that

$$\mathbb{E}\{(\widehat{\psi} - \psi_n)^2\} = \mathbb{E}\{(\widehat{\psi} - \psi)^2\} - \frac{\text{var}(Y^1 - Y^0)}{n} + o_{\mathbb{P}}(1/n)$$

so that the difference-in-means has smaller variance when estimating the sample effect ψ_n. For some intuition, imagine both potential outcomes were observed for each subject: then the sample effect would be estimated without error, but not the population effect.

2.4.3 Difference-in-Means versus Horvitz-Thompson

Note that the difference-in-means estimator is given by

$$\widehat{\psi} = \frac{\mathbb{P}_n(AY)}{\mathbb{P}_n(A)} - \frac{\mathbb{P}_n\{(1-A)Y\}}{\mathbb{P}_n(1-A)} = \mathbb{P}_n\left\{\left(\frac{A}{\widehat{\pi}}\right)Y - \left(\frac{1-A}{1-\widehat{\pi}}\right)Y\right\}$$

which suggests a different version, where we replace the estimated proportion treated $\widehat{\pi}$ with its known population value π:

$$\widehat{\psi}_{ht} = \mathbb{P}_n\left\{\left(\frac{A}{\pi}\right)Y - \left(\frac{1-A}{1-\pi}\right)Y\right\}$$

This estimator is known as the Horvitz-Thompson estimator, hence the *ht* subscript.

Since we are replacing an estimated quantity $\widehat{\pi}$ with its known value π, it seems as if we should gain efficiency. Is this actually true?

It is straightforward to check that the Horvitz-Thompson estimator is also unbiased and consistent; and since it is exactly equal to a sample average, we can apply the CLT to immediately obtain

$$\sqrt{n}(\widehat{\psi}_{ht} - \psi) \rightsquigarrow N\left(0, \operatorname{var}\left\{\left(\frac{A}{\pi} - \frac{1-A}{1-\pi}\right)Y\right\}\right)$$

Now which estimator should we use: difference-in-means or Horvitz-Thompson? Both are unbiased, root-n consistent, and asymptotically normal. Is our intuition correct that it is beneficial to replace the estimate $\widehat{\pi}$ with its known value π? To answer this we will compare asymptotic variances.

Let $\phi = \frac{A}{\pi}(Y-\mu_1) - \frac{1-A}{1-\pi}(Y-\mu_0)$ and $\phi_{ht} = \left(\frac{A}{\pi} - \frac{1-A}{1-\pi}\right)Y$ denote the functions corresponding to the asymptotic variances of $\widehat{\psi}$ and $\widehat{\psi}_{ht}$. Then we have

$$\begin{aligned}\operatorname{var}(\phi_{ht}) &= \operatorname{var}\left(\phi + \frac{A}{\pi}\mu_1 - \frac{1-A}{1-\pi}\mu_0\right) \\ &= \operatorname{var}(\phi) + \operatorname{var}\left(\frac{A}{\pi}\mu_1 - \frac{1-A}{1-\pi}\mu_0\right)\end{aligned}$$

where the last line follows since $\mathbb{E}(\phi \mid A) = 0$ implies that

$$\operatorname{cov}\left(\phi, \frac{A}{\pi}\mu_1 - \frac{1-A}{1-\pi}\mu_0\right) = 0$$

by iterated expectation.

Therefore

$$\operatorname{var}(\phi_{ht}) \geq \operatorname{var}(\phi)$$

and thus the Horvitz-Thompson estimator is *less efficient* than the difference-in-means. This is counterintuitive: here replacing an estimated quantity with its known population counterpart actually reduces efficiency! Usually when we estimate things we get something *less precise* than if we just used the true quantity.

Unfortunately, I do not know of a very satisfying intuitive explanation of this paradox. One way to think about it is as follows: rather than viewing $\widehat{\psi}_{ht}$ as replacing an estimated quantity with a known quantity, one can instead view it as moving away from the sample average $\widehat{\psi} = \widehat{\mu}_1 - \widehat{\mu}_0$ with a noisier version

$$\widehat{\psi}_{ht} = \left(\frac{\widehat{\pi}}{\pi}\right)\widehat{\mu}_1 - \left(\frac{1-\widehat{\pi}}{1-\pi}\right)\widehat{\mu}_0$$

which should degrade performance, merely since sample averages are efficient estimators of means. In other words, the Horvitz-Thompson estimator is using the expected number of treated $n\pi$ rather than the actual number $n\widehat{\pi}$, so that when the actual number differs from its expectation, the averages are not correctly weighted.

Bibliography

P. M. Aronow, D. P. Green, and D. K. Lee. Sharp bounds on the variance in randomized experiments. *The Annals of Statistics*, 42(3):850–871, 2014.

D. D. Boos and L. A. Stefanski. *Essential Statistical Inference: Theory and Methods*. New York: Springer, 2013.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

E. H. Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.

E. H. Kennedy, Z. Ma, M. D. McHugh, and D. S. Small. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B*, 79(4):1229–1245, 2017.

E. H. Kennedy, S. Balakrishnan, and M. G'Sell. Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics (to appear)*, 2019a.

E. H. Kennedy, S. Lorch, and D. S. Small. Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B*, 81(1):121–143, 2019b.

J. Pearl. Causal inference in statistics: an overview. *Statistics Surveys*, 3:96–146, 2009.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge: Cambridge University Press, 2000.